Journal of Multivariate Analysis 128 (2014) 108–119

Contents lists available at ScienceDirect

Journal of Multivariate Analysis

journal homepage: www.elsevier.com/locate/jmva

Hypothesis testing for high-dimensional covariance matrices☆

Weiming Li [a], Yingli Qin [b,*]

[a] Beijing University of Posts and Telecommunications, Beijing, China

[b] Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, ON N2L 3G1, Canada

ARTICLE INFO

Article history:
Received 23 March 2013
Available online 31 March 2014

AMS subject classifications:
62H15
62H10

Keywords:
Covariance matrix
Empirical spectral distribution
High-dimensional
Hypothesis testing
Stieltjes transform

ABSTRACT

This paper discusses the problem of testing for high-dimensional covariance matrices. Tests for an identity matrix and for the equality of two covariance matrices are considered when the data dimension and the sample size are both large. Most importantly, the dimension can be much larger than the sample size. The proposed test statistics are built upon the Stieltjes transform of the spectral distribution of the sample covariance matrix. We prove that the proposed statistics are asymptotically chi-square distributed under the null hypotheses, and normally distributed under the alternative hypotheses. Simulation results show that for finite dimension and sample size the proposed tests outperform some existing methods in various cases.

1. Introduction

Modern statistical analysis often encounters high-dimensional data. For instance, in DNA microarray analysis, the data dimension p (e.g. the number of genes of interest) is typically in the thousands, while the sample size n (e.g. the number of biological samples) is normally in the dozens or at most a couple hundreds. The majority of multivariate statistical procedures are no longer valid since their asymptotic properties are built within a framework where p is fixed and n approaches infinity. Therefore, novel statistical procedures which can handle "large p, large n" or even "large p, small n" need to be developed.

In this paper, we are interested in testing for population covariance matrices when the data dimension p can be much larger than the sample size n. Two tests will be discussed: a one-sample test, testing the identity of a $p \times p$ covariance matrix Σ_p,

$$H_0 : \Sigma_p = I_p \quad vs.\ H_1 : \Sigma_p \neq I_p, \tag{1.1}$$

and a two-sample test for the equality of two $p \times p$ covariance matrices Σ_{1p} and Σ_{2p},

$$H_0 : \Sigma_{1p} = \Sigma_{2p} \quad vs.\ H_1 : \Sigma_{1p} \neq \Sigma_{2p}. \tag{1.2}$$

☆ Weiming Li's research was supported by the Fundamental Research Funds for the Central Universities, No. 2014RC0905. Yingli Qin's research was supported by the University of Waterloo's start-up grant No. 203123.

* Corresponding author.
E-mail addresses: liwm601@gmail.com (W. Li), yingli.qin@uwaterloo.ca (Y. Qin).

http://dx.doi.org/10.1016/j.jmva.2014.03.013
0047-259X/

Classical tests based on the likelihood ratio [1] suffer poor performance when p is not negligible with respect to n, since the sample covariance matrix does not converge to its population counterpart in high-dimensional situations. An important improvement in [2] corrects these likelihood ratio tests to accommodate situations where both p and n can be large but $p < n$. However, this correction cannot be easily extended to situations where $p > n$ as the corrected statistics involve the logarithm of the determinant of the sample covariance matrix, which is singular when $p > n$.

There are a number of works in the literature addressing the testing problems in situations where $p > n$. Ledoit and Wolf [13] modified certain tests, originally proposed by John [11] and Nagao [17], to adapt situations where p and n increase at the same rate. The results were later extended in [5] to the case when p/n tends to infinity or zero. In [23], a likelihood ratio type test was put forward for the identity (and sphericity) test, where only the non-zero sample eigenvalues were included in the likelihood ratio statistic. Schott [20] proposed a test for the equality of several covariance matrices based on the sum of squared differences between elements of the sample covariance matrices. In [24,26,25,9], unbiased estimators of $\mathrm{tr}\Sigma^k/p$ ($k = 1, 2, 3, 4$) were constructed using functions of the sample covariance matrices, from which some tests were developed. In [8,14,7], new estimators of $\mathrm{tr}\Sigma^k/p$ were introduced, and some tests were then investigated based on these estimators.

A common feature among these methods is that they put emphasis on the eigenvalues of sample or population covariance matrices. In particular, the differences in the first few moments of the eigenvalues between the null and the alternative hypotheses. Most recently, [6] presented a two sample test in sparse settings based on the maximum of standardized differences between the entries of the sample covariances. For the study of the asymptotic properties of relevant tests, one is referred to [24,8,14,18].

In this paper, we investigate the testing problems (1.1) and (1.2) from random matrix theory point of view. We focus on the empirical spectral distributions rather than the moments of population (or sample) eigenvalues. By looking at the empirical spectral distributions, we are able to utilize the complete distribution of all eigenvalues. One may see that the new tests can detect a small shift of the bulk eigenvalues. Moreover, the power of the tests increases as the dimension p increases.

The success of this strategy relies on the convergence theorem of empirical spectral distributions in random matrix theory. The *spectral distribution* (SD) G^A of an $m \times m$ Hermitian matrix (or symmetric in real case) A is the measure generated by the eigenvalues $\{\lambda_j\}$ of A,

$$G^A = \frac{1}{m}\sum_{j=1}^{m}\delta_{\lambda_j},$$

where δ_b denotes the Dirac point measure at b. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be a sequence of i.i.d. zero-mean random vectors in $\mathbb{R}^p$ or $\mathbb{C}^p$, with a common population covariance matrix Σ_p. The sample covariance matrix takes the form of $S_n = \sum_{k=1}^{n}\mathbf{x}_k\mathbf{x}_k^*/n$, where $\mathbf{x}_k^*$ stands for the conjugate transpose of $\mathbf{x}_k$. We are interested in the limiting relationship between the following two SDs as both $p \to \infty$ and $n \to \infty$:

$$H_p := G^{\Sigma_p} \quad \text{and} \quad F_n := G^{S_n},$$

which are referred as *population spectral distribution* (PSD) and *empirical spectral distribution* (ESD), respectively.

Following the conventional assumption in random matrix theory, we assume that H_p weakly converges to a limiting distribution H, as $p \to \infty$. Then under some regularity conditions, as p, n both tend to infinity with $p/n \to c > 0$, the ESD F_n converges to a non-random distribution F which relates H via the Marčenko–Pastur (MP) equation through Stieltjes transform, see (2.1). Particularly, if $H = \delta_1$ then the distribution F is the MP law.

The properties of the limiting distribution F offer us a new way to test for population covariances matrices. To test the hypothesis (1.1), we propose to measure the difference between the ESD F_n and the MP law; to test the hypothesis (1.2), we propose to measure the difference between the two ESDs.

The rest of this paper is organized as follows. In the next section, we discuss the test for the identity covariance matrix (1.1). The proposed test statistic is extended to testing for the equality of two covariance matrices (1.2) in Section 3. Section 4 reports simulation results. Conclusions and remarks are presented in Section 5, and proofs are postponed to Appendix.

2. Test for the identity of Σ_p

2.1. Main assumptions and the Marčenko–Pastur equation

Our model assumptions are as follows.

Assumption (a). Both $n, p \to \infty$ with $c_n = p/n \to c \in (0, \infty)$.

Assumption (b). There is a doubly infinite array of i.i.d. random variables $(w_{jk}), j, k \geq 1$ satisfying

$$\mathbb{E}(w_{11}) = 0, \qquad \mathbb{E}(|w_{11}|^2) = 1, \qquad \mathbb{E}(|w_{11}|^4) < \infty,$$

such that for each pair of (p, n), let $W = (w_{jk})_{1\leq j\leq p, 1\leq k\leq n}$, the observation vectors can be represented as $\mathbf{x}_k = \Sigma_p^{1/2}w_{\cdot k}$ where $w_{\cdot k} = (w_{jk})_{1\leq j\leq p}$ denotes the k-th column of W and $\Sigma_p^{1/2}$ stands for any Hermitian square root of Σ_p.